7-27-2021

An Exploratory Assessment of Peer-reviewed Life Science Literature Statement Readability Used in Text-based Interpretive Analyses

Jeremy Jackson
jsj3398@rit.edu

RIT

An Exploratory Assessment of Peer-reviewed Life Science Literature Statement Readability Used in Text-based Interpretive Analyses

by
Jeremy Jackson

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics

Committee: Dr. Gary Skuse, Dr. Gordon Broderick, and Dr. Matt Morris

School/Department of Thomas H. Gosnell School of Life Sciences

College of Science

Rochester Institute of Technology

Rochester, NY 14623-5603

July 27, 2021

Abstract

This project looked to explore the process of one of the scientific community's leading text mining software, Pathway Studio, which efficiently streamlines the tedious task of information sorting and gathering in a clinical setting. To do this Pathway Studio implements a sentiment scoring algorithm which decides, based on interpreting literature similarly to natural human understanding, what sections of literature are most relevant to a given search request and provides corresponding peer-reviewed work in the industry. The novel statistics derived of its performance are used to establish an average form of a statement which reflects that of a standard level of human understanding of text. This was done by determining a mean number of words found between a given source and target, and of those words how many occurring are verbs. This was found to be 10 words to 1 verb occurring on average within this project's dataset. In addition, this project sets the foundation and proposing a new readability scoring formula that adds insight to the structure of a citation and how it may be interpreted relative to a researcher's average level of understanding of the scientific community's peer-reviewed literature. The structures observed were scored anywhere from 0 to 10, where 0 represented unreadable citations, 1 represented a citation containing the mean number of words and verbs between its source and target, and 10 being representing a citation that was verb-rich, regardless of the number of total words found between its source and target. According to this readability scoring formula it was found that the variation of the observed citation used for this project from the determined mean citation structure was 70.5%.

Introduction

Natural Language Processing (NLP) is a method of analyzing natural human language through machine learning to interpret text as it pertains to parts of speech (POS), speech patterns, sentiment, language rules, and grammar as directly indicated by the writer (Nadkarni et al., 2011). NLP is a branch of artificial intelligence originated from a Swiss linguistics philosopher named Ferdinand de Saussure and his successors Albert Sechehaye and Charles Bally (Stoltz, 2018) who paved the way for its development throughout the 1940s through the development of the Structuralist Approach, which encompasses the foundation for teaching the English language through grammatical structures, speaking rules, rules of comprehension, and grammar mechanics in speaking, reading, and writing (Saussure & De, 1959).

Studying NLP has been an interest since approximately the 1950s (Nadkarni et al., 2011) and can be seen today in simple forms such as e-mail filters, translators, search engines such as Google, and predictive text and virtual assistance on handheld devices. This is achieved through what is called machine learning, a modeling procedure based on the artificial intelligence used so that machines/algorithms can learn from experience to make 'informed' choices. There are three methods; supervised, unsupervised, and reinforcement (Alloghani et al., 2020; Sarker, 2021)

NLP through ML is currently being used as an asset in text mining/analytics for published literature to accelerate the pace at which information can be gathered for research purposes to aid medical advancement. This becomes especially true for new threats to our society such as the COVID-19 pandemic for which quick research was necessary.

Outside of direct application to clinical data collection there is also something to be said about the cumulative collection of clinical information being important to knowing how to move forward. Pathway Studio (PS), developed by Elsevier, is currently the industry standard for text

mining and analysis within the life sciences and aims to streamline the tedious process of searching through the literature for prospective hypotheses and methodologies (Nikitin et al., 2003; Sharp, n.d.). There are several variables that need to be considered when gathering research material such as authorship and journal credibility, type of relationship between the subject and research question, context of research interest in a paper, and methods to be used. When handling so many variables it becomes difficult to gather the desired amount of literature from various sources because confirming their relevance is very time consuming. Pathway Studio seeks to reduce this task by performing this search, returning the results, reporting where it was found per source and target provided by the user.

The vast number of challenges in NLP research today can be broken down into a few overarching categories; ambiguity, synonyms, domain specific language, and low resource languages (Khurana et al., 2017). Ambiguity is part of the root to NLP research, having machines interpret text as humans would. Ambiguity refers to the ability of text to be understood in multiple ways and can be dissected into 4 different sub-categories; lexical ambiguity, syntactic ambiguity, semantic ambiguity, and discourse (Anjali & Babu, 2014). Lexical ambiguity refers to the ambiguity in single word homonyms. Syntactic ambiguity now focuses on an overall phase where qualifiers are not directly attached to a subject and therefore can be interpreted as belonging to one or more subjects. Sematic ambiguity is where the overall interpretation of a sentence can have multiple meaning. This is a combination of the previous two types. The last type of ambiguity is discourse which covers the misinterpretation of a sentence based on omitting information while still speaking with regards to said information just because it was present in a previous sentence, therefore changing the meaning of a sentence when it stands

alone. To work around this one option is to make use of working memory model (Adams et al., 2018).

Domain specific language is another challenge as every model must be specific to a certain area of research. A model proposed for clinical trials for example cannot be used for electrical engineering material because of the necessary lexicon and various colloquialisms introduced in differing quantities within each area. Low resource languages fall under a similar challenge where technological advancement tends to only assess problems within the language it was created. One of the steps towards working around this challenger is LASER (Language-Agnostic Sentence Representations) by Facebook released in January 2019 (Artetxe & Schwenk, 2019). This uses zero-shot transfer for NLP models to score text in one language, to then translate to another to accommodate multilanguage sentiment scoring.

A combination of these text processing challenges also introduces an additional challenge for POS tagging because this is based in ambiguity. The harder it is to discern what responsibility a word takes on in a sentence, the harder it is to give it one exact POS tag. This may be overcome by incorporating the probability and the model used. This means taking into consideration the most likely POS tag of a word in coordination with the type of model used, i.e. one preferring nouns or verbs (Roy & Purkayastha, 2016).

This project ties into these challenges by addressing ambiguity in relation to the novel statistics it seeks to produce by filtering the dataset to exclude citations that contain too much ambiguity making it impossible to pinpoint either an exact start, end, or both. This, therefore, helped to increase the credibility of the derived statistics. This project also addresses the use of POS tagging by using the Apache openNLP package to tokenize citations. Tokenizing involves splitting citations into smaller units, in this case splitting sentences into their individual work and

punctuation components (*Tokenization as the Initial Phase in NLP*, 1992). This progress was paired together with filtering to exclude incorrect tagging of punctuation where necessary.

This project aims to derive novel statistics which will provide additional insights into the output of PS to know to what degree citations are relevant to their given sources and targets. This was done through establishing an average citation structure from the PS output which in turn gives rise to understanding how we tend to communicate concepts through peer-reviewed literature. The importance of this project stems from the need to quickly sift through data with a relatively large confidence in reported results. One way to do this is by assessing the readability given material in accordance with the suggested sixth grade reading level used as a universal standard. This is important based on the dispersion of reading level among the general population in the US. Most adults average at a reading level that of an eight-grade level, while overall, 20% of the population reads at/below a fifth-grade level (Safeer & Keenan, 2005). This greatly attributes to the importance of having healthcare literature be established on a standard to accommodate understanding, by as much of the population as possible. Recent advancement in NLP has shown the need to have a standard create through a search and characterization of results done based on 6 different readability indexes. It was found that the average reading level found among different types of healthcare material pertaining to Achilles rupture and reconstruction was 10.7 ± 2.54 (Perez et al., 2020) One of the six scoring algorithms used during this study is the Flesch-Kincaid grade level which according to a score of 10.7, between 10.0-30.0 score , indicates a reading level very difficult that is best understood by university graduates (Kher et al., 2017). This is significantly higher than the average reading level of the population hence being material that is widely inaccessible to the public in terms of utilizing in accordance with their abilities. This is an instance showing that the readability of text within life science

literature is declining approached 2015, and indicative of continuing beyond to modern years of research (Plavén-Sigray et al., 2017). This project approaches the readability assessment like a simplified version of this algorithm to assess the readability on a smaller scale i.e., individual statements. As modern research began to produce more information, it became more important to gather and store said information therefore leading to a greater need to rely on teaching machines to interpret as it makes this possible due to its much quicker capabilities.

All statistics are derived using R due to the method chosen to analyze the text. Once citations are read in, they are quickly organized into data frames for quick access and different elements referred to by index positions or by association through other integer values. R stores data in physical memory which can pose a problem for handling big data however, it comes with a base collection and a wide range of importable packages that make its exploration and development easier.

The major packages used for this project include openNLPmodels.en and openNLP for fast analysis of different text components. The openNLP package is a machine learning-based library developed and maintained by the Apache Software Foundation that is used for processing text. This comes in many forms for tokenizing, sentence segmentation/chunking/parsing, and POS-tagging (*Apache OpenNLP Developer Documentation*, 2021). This project specifically takes advantage of the POS-tagging. To perform POS-tagging the *annotation()* function of this package is used to tokenize a citation, which includes punctuation and special characters. The function takes in the current vector that was tokenized and uses its maxent_POS_Tag_Annottaor to generate each element's respective part of speech tagging (*Maxent_POS_Tag_Annotator Function - RDocumentation*, n.d.). The other NLP package, openNLPmodels.en was used in conjunction as a token annotator and string manipulator developed by the Institute for Statistics

and Mathematics and the Research Institute for Computational Methods. This was used in order to produce the correct annotations for every single element and to easily parse them together to form a human readable format (*Datacube Resource Homepage*, 2006).

Other packages used alongside base R were tidyverse and ggplot2. Tidyverse was imported to utilized the *between()* function that was used to quickly grab elements given a start (source) and end (target) location instead of having to use slower methods of manually specifying stepping through a vector one index at a time (*Do Values in a Numeric Vector Fall in Specified Range?,* n.d.). The package ggplot2 was used to visualize several different statistical features, including the final categorization and interpretation of citations using the newly proposed sentiment scoring formula (figure 11)

Research Question

Can production of novel statistics with this software allow for proposing a readability scoring formula that gives more insight int the returned citation quality? What is observed to be the most acceptable structure of statements that closely reflects human interpretation?

Methods

Data origin/curation

The data comes from the results of a coronavirus cytokine storm model based on Morris et al. 2019, represented by the immune regulatory circuit (Figure 1) In Figure 1 the relationships between the different targets are noted as positive, negative, and the degree of relationship estimated by each citation. The source of this model is derived from a collection of citations, a

representative sample is provided in Figure 2, for which each citation record has a source, target and a relationship type.

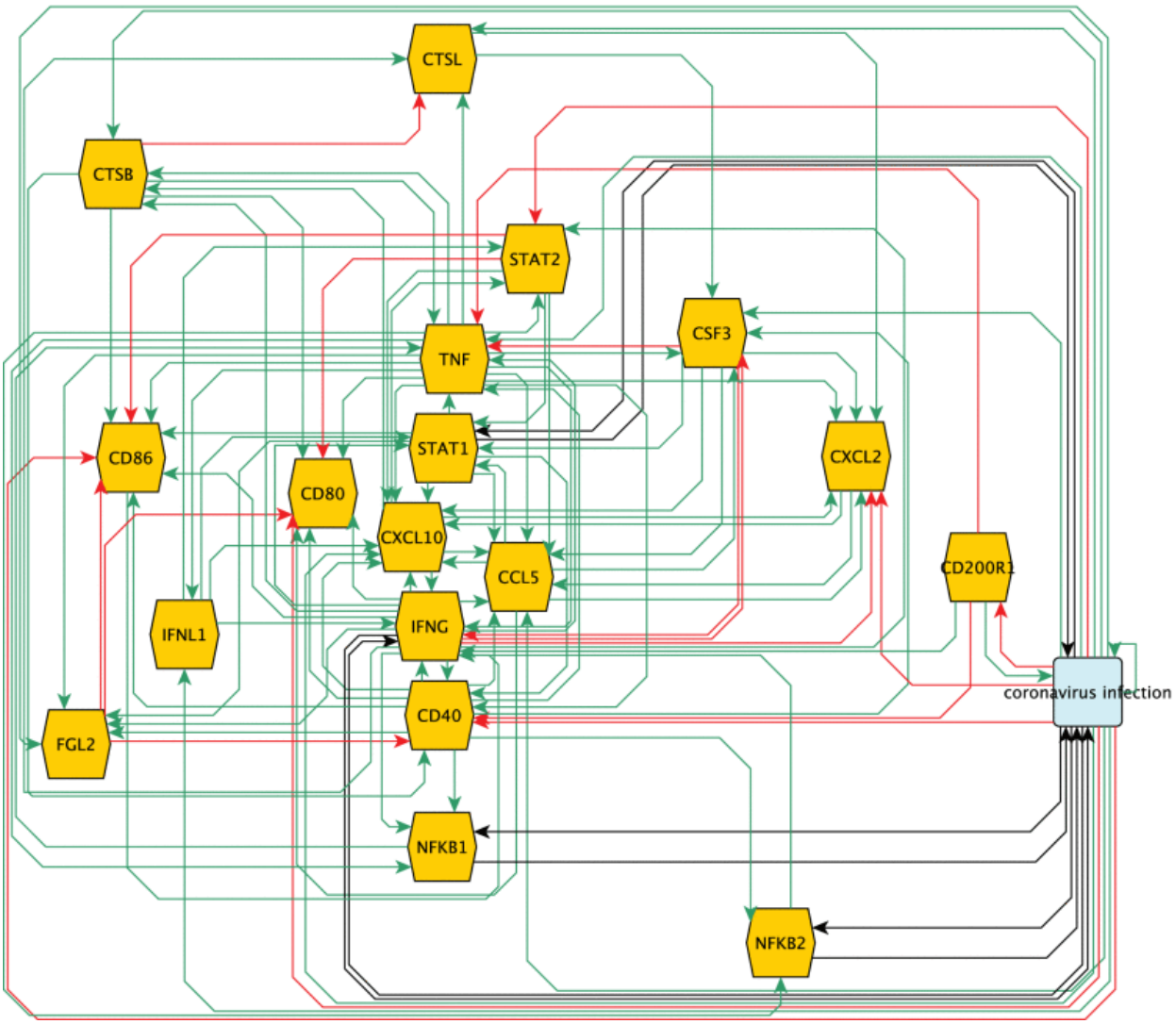

Figure 1. Immune regulation circuit assembled from 2,653 journal publications passed through Pathway Studio (Morris et al., 2020).

Source	Entity1 Object Type	Relation Object Type	Relation Effect	RefCount	Relation PMID	Relation Sentence	Target	Entity2 Object Type
STAT2	Protein	DirectRegulation	positive	135	18579593;1744	Here we demonstrate that although STAT1 interference results from protein in	STAT1	Protein
FGL2	Protein	Expression	negative	1	31409352	Mechanistically, we found that Soluble FGL2 hindered the expression of major	CD80	Protein
TNF	Protein	Expression	positive	54	27773519;2847	There are a number of reports describing a correlation between mitogen-activ	CD86	Protein
IFNG	Protein	Expression	positive	64	27502468;2750	IL-22 also suppresses IFN-Î³-induced expression of MHC class I, MHC class II, I	CCL5	Protein
TNF	Protein	Expression	positive	66	1469;29126927	While various pro-inflammatory stimuli including LPS, TNF-Î±, interleukin-1Î², a	CSF3	Protein
CD80	Protein	Expression	positive	44	27717695;1562	CD80 is known to induce Th1 responses, leading to IFN-Î³ and IL-2 production,	IFNG	Protein
CSF3	Protein	Expression	negative	55	28756227;9528	Both doses of G-CSF significantly increased IL-6 and TNF-Î± mRNA levels at We	TNF	Protein

Figure 2. Snippet of the Pathway Studio output which shows the subject (source) and query (target) relationship attributes. The total number of records is 128 with a total of 2,653 citations between them all (not proportionally distributed).

To begin, each record was deconstructed into individual citations by establishing suitable delimiters. Most source-target pairs have many citations supporting them, delimited by a semi-colon. This posed difficulty considering that semicolons are a regular component of the literature and therefore specific iterations of the semicolon groupings were used to identify them as delimiters. These iterations include "._;", ".;", "._;_", and ".;_" (underscores represent spaces in the text). These iterations would signify the end of a citation by the period, and the beginning of a citation by a semicolon and sometimes another period. All citations were then split into the following categories, as in Table 1:

A – subject and query appear verbatim (at least once each).

B – subject only appears verbatim (at least once).

C – query only appears verbatim (at least once).

D – neither subject nor query appears verbatim.

This project only focuses on category A until the very last step because it is the only category in which both a source and target appear verbatim. Without these both it would be impossible consistently extract words between them when the position of one or both in unknown. So, to give the findings more meaning, by knowing exactly what citations are being, only category A is used for the analyses up until the last step of comparing an average readability score in each. The first method tried involved cutting out citations in categories B, C and D that did not have a

verbatim occurrence of both source and target. However, this posed a problem when it came to supporting the findings with a substantial sample group. This would not allow for at least the same number of citations used in category A.

Table 1. Example sentences from the Pathway Studio output that reflects the different categories of data this project splits it into. Category A is subject and query appearance verbatim at least once each, category B is subject only verbatim appearance at least once, category C is query only verbatim appearance at least once, and category D is neither appearance. Each example presented is the first occurrence from the PS output. Sources are highlighted in green, and targets are highlighted in orange for easy identification in the supporting citation.

Category	**Source**	**Target**	**Supporting citation** (presence of source and target highlighted)
A	STAT2	STAT1	Here we demonstrate that although STAT1 interference results from protein interactions within a V protein N-terminal region encompassed by amino acids 110 to 130, detection of STAT1 interaction and IFN-gamma signaling inhibition requires the presence of cellular STAT2.
B	STAT2	STAT1	Then signal transducer and activator of transcription1Î± binds to STAT2 and also becomes phosphorylated
C	STAT2	STAT1	This can potentially lead to an uncontrolled rise of STATs levels and activity, however this positive-feedback loop is controlled by STAT1/2-inducible SOCS1, which inhibits IFNAR receptors and attenuates STAT1/2 phosphorylation35,36.
D	STAT2	STAT1	Phosphorylated STAT-1 binds STAT-2 and p48 to form the ISGF3 complex, which translocates to the nucleus.

It is important to note that case sensitivity was ignored for this project. From here the project focuses on category A to establish an environment which has the least unaccountable

variables while performing text analytics. This is important because at a time when there is no verbatim occurrence of a subject or query, meaning it may be present in the citations but possibly just recognized by a synonym, abbreviation, acronym, etc.

POS statistical analysis

The second step of this project begins working on category A of the dataset. To begin *unlist(strsplit()*) was used to separate each citation into individual words, with each word being assigned its own index. This determined the position of the source and target within a citation. If it was the case that the source occurred once but the target occurred multiple times, then the closest instance of the target, by index position to the source, was chosen as the target boundary. If *vice versa* then the closest instance of the source was paired with the only instance of the target. In the case that both occurred multiple times then the source-target pair used would be that which had the smallest number of words in between. This methodology considered the best case, where one word and/or verb in between can be seen as one of the most direct ways a relationship may occur.

The average number of words was determined by using *grep()* to find the position of both the source and target, among all citations, the average number or words was found by

Sum((source index – target index) -1) / (no. of citations)

The average number of verbs was found by POS tagging every element of each citation using *annotate()*. The output of this was interpreted as plain text by adding an abbreviation to the end of each element describing what they are. Those tagged with VB (verb), VBG (verb gerund), VBD (verb past tense), or VBP (verb, present tense not 3rd person singular), in between the known indices of the source and the target were counted towards the verb count in a citation

(Santorini, 1990). The average verb count hereafter was found similarly to the average word count, by

sum(no. verbs between source index & target index) / (no. of citations)

These statistics go on to define what is the most acceptable composition between a subject and query to consider a citation as containing a relationship between subject and query, A0 (average level 0). Hereafter a proportion about this point was established to indicate the percentage of citations that fell within its standard deviation to be considered an acceptable citation returned, against the total number of citations returned by Pathway Studio. To do this *stat_ellipse()* was used to depict the relationship as seen in figures 9 and 10. This is the confidence interval around A0 that acceptable citations fell within. This showed the approximate percentage of citations within category A that would are acceptable as an understandable citation because those points fell within this ellipse with A0 at its center.

It is important to do this step to form the basis of our analysis whereby obtaining the observed average citation format to establish a base value for the later proposed readability scoring formula in Figure 11.

Readability scoring exploration

Pathway Studio on its own already produces its results based on a sentiment score for each citation found however, to gain more insight into the nature of the relationship between a subject and query, this project proposed an additional readability scoring formula that defines the status of a citation with respect to the average composition of a relationship established in step 02 above. Much like a Z-score which defines a raw score about a mean, this formula shows the

score about this average which will indicate the nature of the verb and word count compared to A0.

The next and last step of this project was to apply the analyses so far to the remaining categories of the dataset. This would then provide a comparison to the found readability score of category A to the remainder. This is done because category A is the only one with a verbatim appearance of both source and target. After this the dataset for each category was expanded to apply the formula to a wider range of possible citations by using any possible pairing in each. This would later give rise to the justification of the average citation structure and visualization of the null distribution done.

Results

The first step of this project was to narrow down the dataset to adhere to a specific set of variables. The data captured in citations that include both their given verbatim source and verbatim target, appearing at least once, namely those in category A. The entire dataset is recognized as having source and targets occurring at least once however, some sources and targets do not occur in their given form but instead as synonyms, acronyms, abbreviations, or symbols. To carry out this project with all these possibilities, a vast lexicon that covers all possible alternatives would be necessary and it would have to be specific to this field. The target subset of data for this project includes verbatim occurrences of both source and target, as least once within each citation (Figure 1).

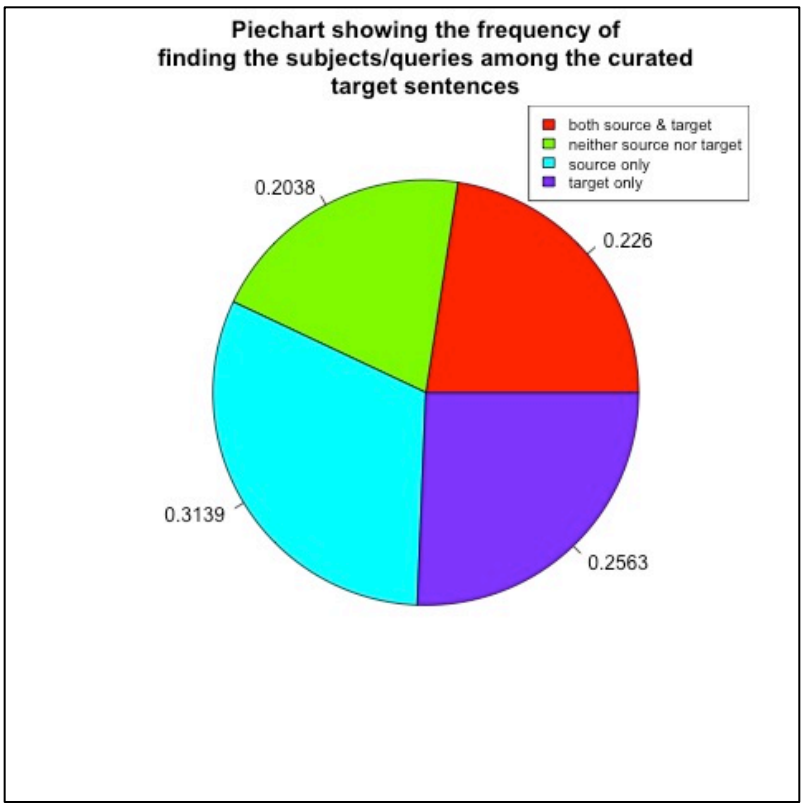

Figure 3. The distribution of categories A, Bs, C, and D. A corresponds to those citation in which the source and target appear verbatim (at least once each), B includes citations in which the source only appears verbatim (at least once), C includes those in which the target only appears verbatim (at least once), and D includes those in which neither source nor the target appear verbatim.

Figure 3 shows category A, the portion of the overall dataset this project focuses on, makes up approximately one quarter of the entire dataset (i.e. 22.60%). Category B, C, and D make up 31.39%, 25.63%, and 20.38% respectively of the overall dataset.

From here the average number of words between a given source and target was found. In Figure 4 below, each citation was denoted as belonging to a specific source-target pairing which resulted in 48 unique pairings, each of which had a disproportionate number of citations which support them. After considering all of category A's citations it was found that the average number of words found between each source-target pairing was 9.570.

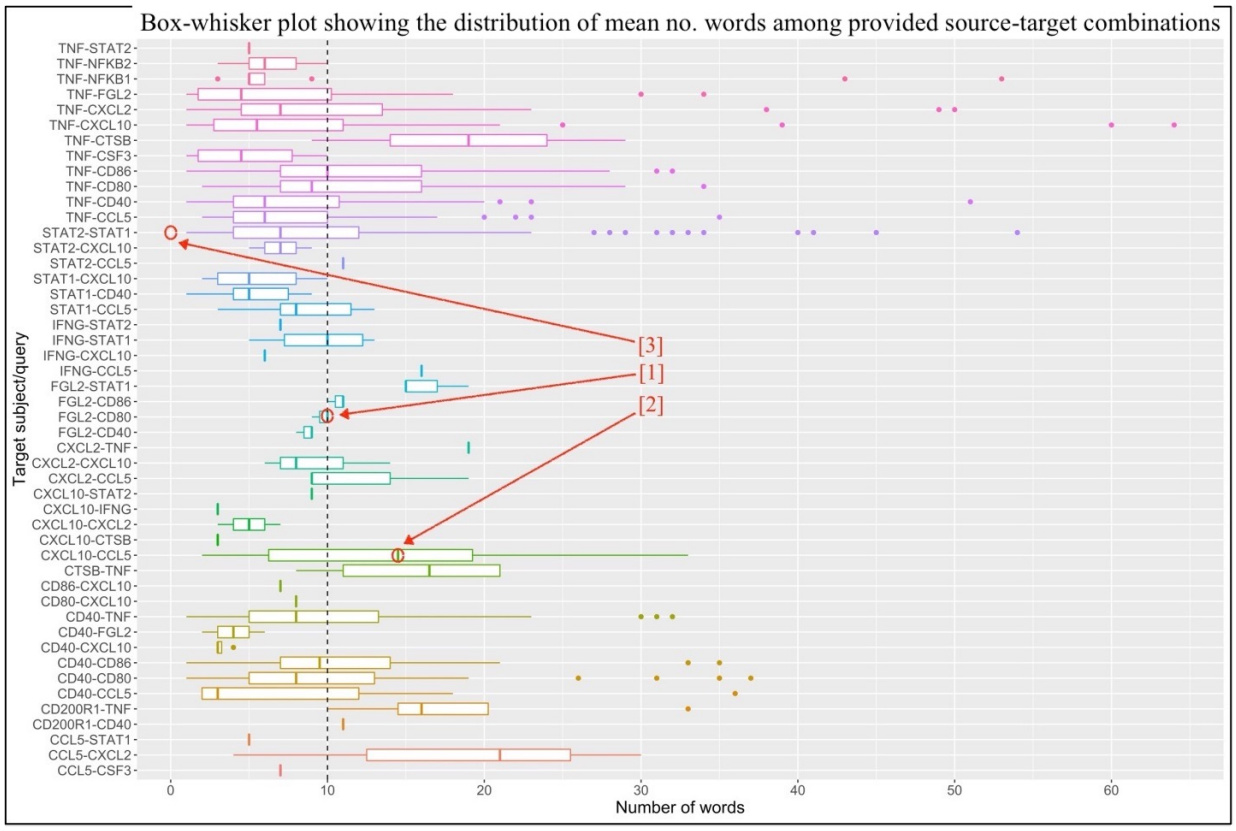

Figure 4. The average number of words found for each of the 48 unique subject-query pairings. Positions indicated by [1], [2], and [3] are citations in which there are different word counts between source and target. These are further explored in Figures 5 and 6. The vertical dashed line at 10 represents the observed average number of words found between a source and target of category A citations.

In Figure 4, there are 3 points indicated with red arrows. These are specific means chosen that [1] represent a citation found holding the same number of words between a source and target as the mean, [2] showing above the mean number of words at 14, and [3] showing below the mean at 1 word. Figure 5 shows this expansion where you can see the citations at that location and their composition. Panel A shows [1] where the number of words is found to be 10 between the source and one target. This is observed to be the most acceptable level of understanding according to the previously found mean word length between pairings. This project's observed

mean number of words falls near what is observed in other literature, that being 12-17 (Deveci, 2019).

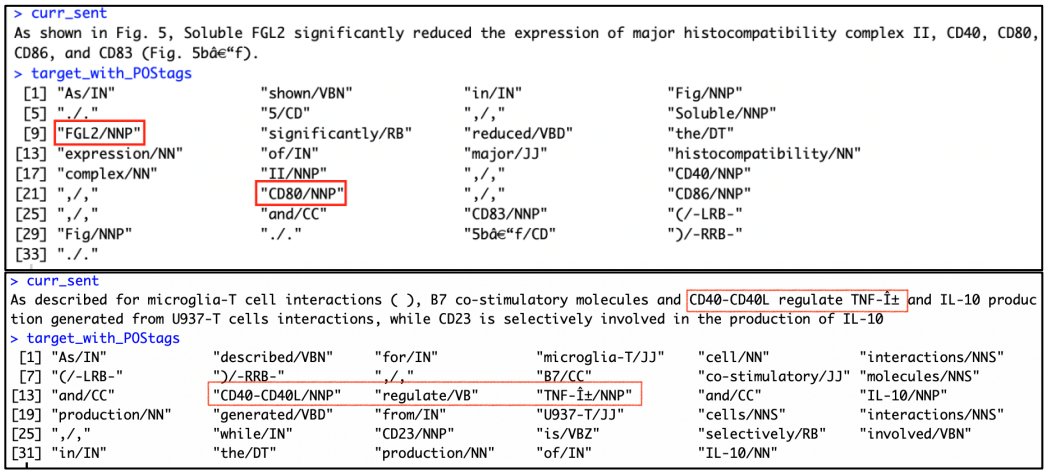

Figure 5. Citation [1] indicated on Figure 4, where the word count between subject and query is at the average of 10 in the first panel. The second panel shows an example of a citation which only has one degree of separation between subject and query as well as the intervening word being a verb. This is seen as an alternative to the most understandable citation format. The panels begin by showing the 'curr_sent', current sentence being looked at, followed by the sentence broken down into its POS tagging for each element.

Panel B in Figure 5 shows another instance in which panel A can be compared as an alternative acceptable citation where there is only one degree of separation between the source and target, specifically a verb. This is at the same level of acceptability because the fact that the only word linking the pair is a verb indicating direct relationship between the source and target.

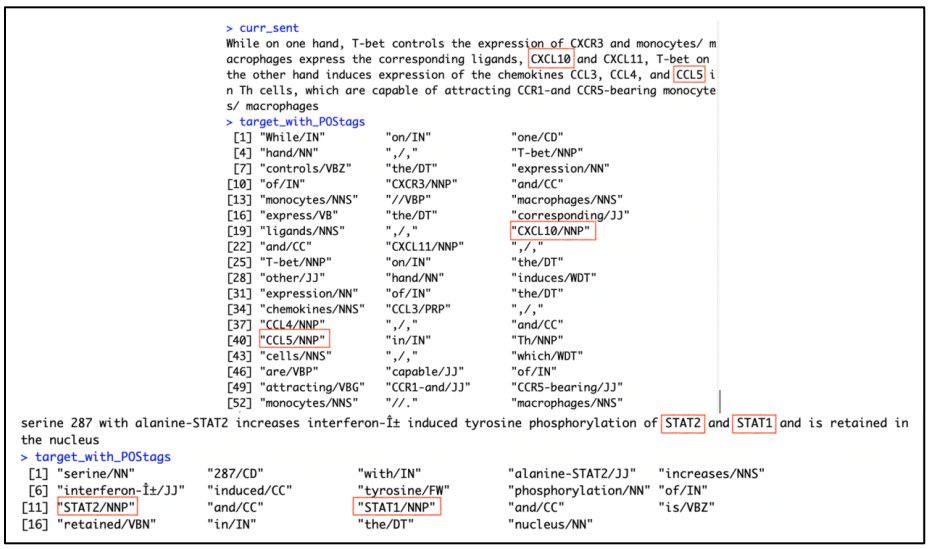

Figure 6. Citations [2] and [3] as indicated in Figure 4 that represent two deviations away from the average citation format. The first panel shows 15 words separating the source and target (i.e. higher than the average) and the second panel shows just one word separating the source and tarbet but it is not a verb. The panels show the 'curr_sent' or current sentence being looked at, followed by the sentence broken down into its POS tagging for each element.

Like Figure 4, Figure 7 above shows the average verb count between each unique source and target. This mean was found to be 0.8935574. Figure 8 shows example citations with respect to calculating the average verb count. Panel A shows the case which has an above average verb count between its source and target, panel B shows the average of 1 verb, and panel C shows below the average with 0 verbs occurring. Note that Panel C also has adjectives denoted by "JJ" because these may be interpreted as indicative of a relationship established between a source and its target.

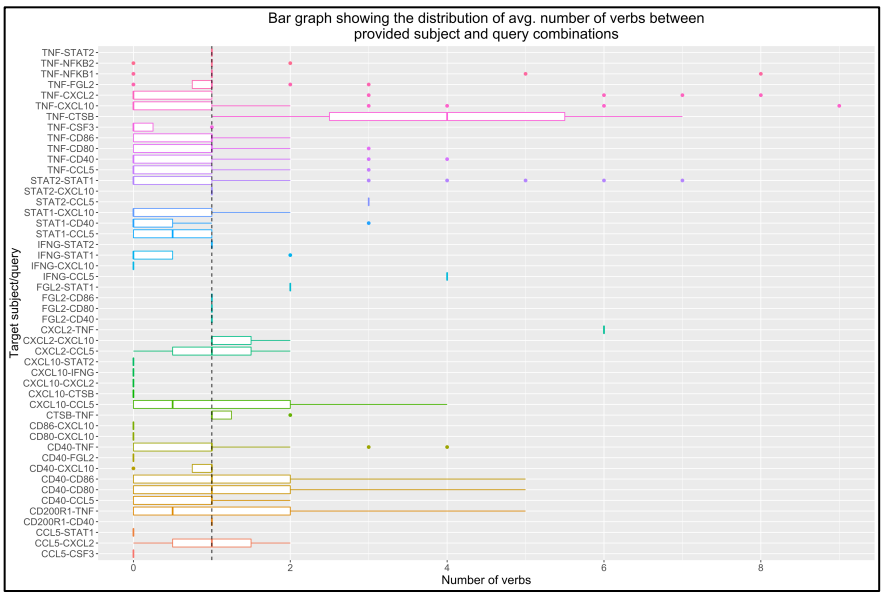

Figure 7. The average verb count between 48 unique source -target pairings. The vertical dashed line at 1 represents the observed average number of verbs found between a source and target among category A citations.

```
> curr_sent
TNF attenuated the NIK:IKK1 activity and concurrently induced the expression of Nfkb2 mRNA thro
ugh the canonical pathway; these together accumulated unprocessed p100 in LTÎ²R-stimulated cell
s
> target_with_POStags
 [1] "TNF/NNP"              "attenuated/VBD"       "the/DT"               "NIK:IKK1/NNP"
 [5] "activity/NN"          "and/CC"               "concurrently/RB"      "induced/VBD"
 [9] "the/DT"               "expression/NN"        "of/IN"                "Nfkb2/NNP"
[13] "mRNA/NNS"             "through/IN"           "the/DT"               "canonical/JJ"
[17] "pathway/NN"           ";/:"                  "these/DT"             "together/RB"
[21] "accumulated/VBN"      "unprocessed/JJ"       "p100/NN"              "in/IN"
[25] "LTÎ²R-stimulated/JJ" "cells/NNS"
> curr_sent
In addition, TNF-Î±, but not interleukin-1Î², enhanced the IFN-Î³-induced production of Mig/CXC
L9 and IP-10/CXCL10 in eosinophils
> target_with_POStags
 [1] "In/IN"                "addition/NN"          ",/,"                  "TNF-Î±/NNP"
 [5] ",/,"                  "but/CC"               "not/RB"               "interleukin-1Î²/JJ"
 [9] ",/,"                  "enhanced/VBN"         "the/DT"               "IFN-Î³-induced/JJ"
[13] "production/NN"        "of/IN"                "Mig/CXCL9/NNP"        "and/CC"
[17] "IP-10/CXCL10/NNP"     "in/IN"                "eosinophils/NNS"
> curr_sent
CD40 ligation results in marked p44/42 mitogen-activated protein kinase activity and TNF -Î± pr
oduction in microglia deficient for CD45
> target_with_POStags
 [1] "CD40/IN"              "ligation/NN"          "results/NNS"
 [4] "in/IN"                "marked/JJ"            "p44/42/NN"
 [7] "mitogen-activated/JJ" "protein/NN"           "kinase/NN"
[10] "activity/NN"          "and/CC"               "TNF/NN"
[13] "-Î±/JJ"               "production/NN"        "in/IN"
[16] "microglia/NN"         "deficient/NN"         "for/IN"
[19] "CD45/VBG"
```

Figure 8. Example citations indicating verbs found between source-target pairs. The panels begin by showing the current sentence being looked at, followed by the sentence broken down into its POS tagging for each element. Red boxes represent sources and targets while blue boxes represent verbs observed in between them.

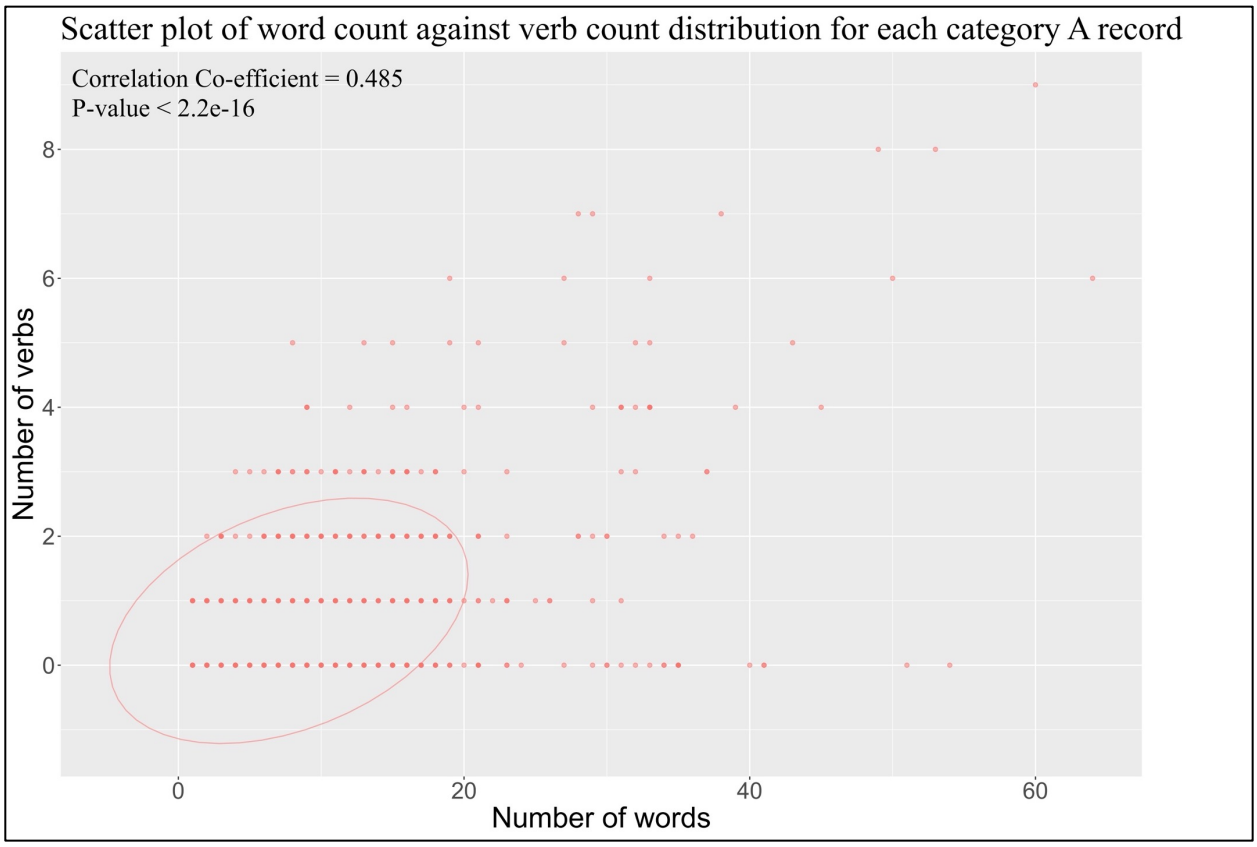

Figure 9. Scatter plot showing the distribution of word VS verb count for all 1123 citations found in category A of the dataset. Each point represents a single citation but as it becomes darker this indicates multiple citations fall under the same verb/word distribution. Darker points indicate overlapping data points. The ellipse seen in the lower left quadrant of the graph is a data ellipse which shows the standard deviation about the observed average number of words and number of verbs for category A citations. This relationship displayed was found to have a significant correlation of 0.0485

Once the average word and verb counts were found, the project moved onto using these to draw additional insights. The first of which was the distribution of citations according to their verb count to word count ratios. Figure 9 shows a distribution of the total 1123 citations in category A. Considering the verb count as a subset of the word count, if the word count is 0, then the verb count must be zero as well. Each point represents a citation's verb to word count ratio distributed across the data set. The red ellipse visually represents the standard deviation about the

average. Ones that fall within the ellipse are deemed to be within an acceptable range about the average level of understanding, that a citation composed of on average 10 words containing 1 verb, between the source and target (Friendly et al., 2013).

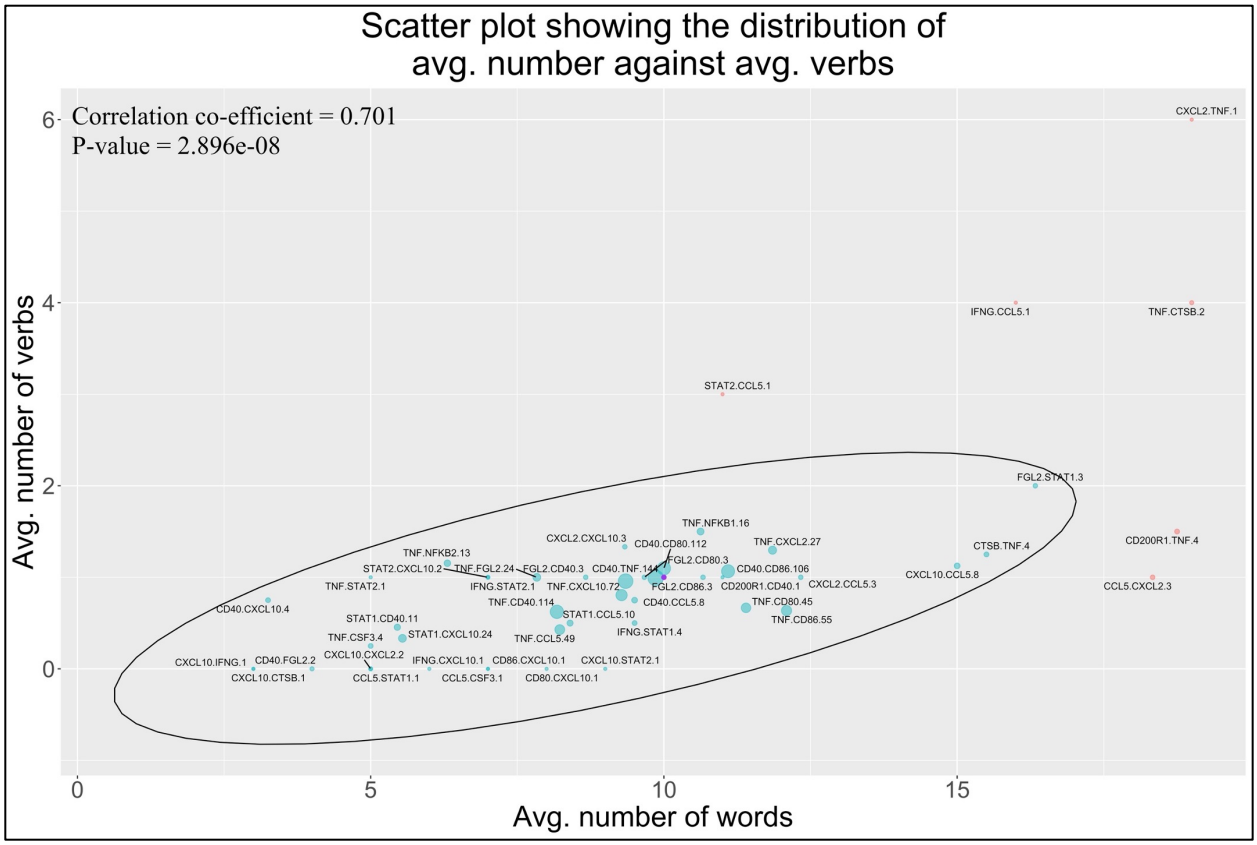

Figure 10. The distribution of the average number of words against the average number of verbs found between a given source and target. Larger data points indicate a higher number of citations that support the proposed source-target pairing. The ellipse in the lower half of the graph represents the standard deviation about the observed average number of words and number of verbs for category A citations among each unique source-target pairing.

Having seen the distribution in Figure 9, the 1123 citations were binned to their respective source-target pairings and the average verb count and word count ratios per unique pairings were plotted in Figure 10. Like Figure 9, a data ellipse was drawn to visualize the

threshold of acceptable citation about the average point (in purple). In this case it was found that 87.5% of the unique citations were seen as acceptable according to the established average.

The final portion of the project establishes a sentiment scoring formula that gives insight as to the degree of difference from the mean, i.e., the most acceptable level of understanding noted by a citation comprised of 10 words (word constant), one of which is a verb (verb constant). It is important to note that the formula shown in Figure 11 must only be used when the word count is 1 or higher because when it is zero, the formula is unable to produce a valid result since it will be dividing by zero.

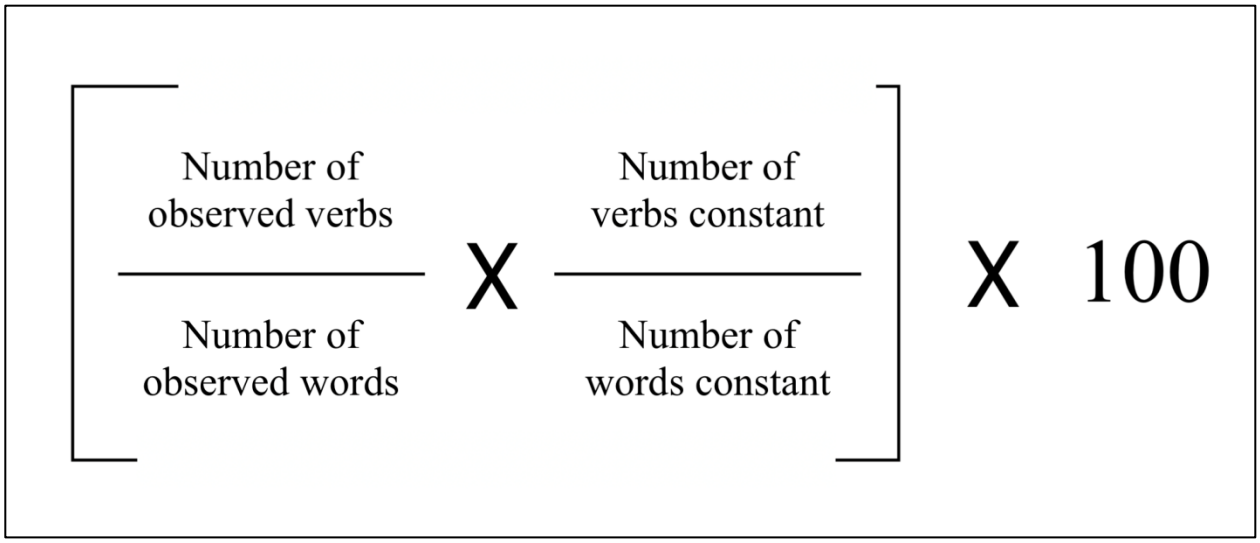

Figure 11. Readability scoring formula proposed by this project as a ratio of observed verbs to verbs constant against the observed word count to words constant.

To demonstrate the proposed formula, Table 2 shows a series of unrelated sentences with the formula applied. The first sentence is a representation of the average case, A0, scoring a readability score of 1. Below this, as the number of words vary, and number of verbs vary the score begins to fluctuate depending on which variable is changed. As the readability score that

this formula produces hovers around 1.0, it means that the citation has a word-to-verb ratio close to 1.0. As it approaches 10 this means that out of all the words that are located between a source and target, almost all are tagged as verbs. Once the sentiment score is 10 this means that all the words found between a source and target are verbs. As the sentiment score approaches 0, this represents a citation that has decreasing number of words, which also means 1 or no verbs are founds either.

Table 2. Table showing the sentiment score above being applied to regular statements. The subject and target are highlighted in yellow, which in this case it does not matter which is which, and the green highlights represent verbs (b). Words that are not highlighted between the subject and target are counted towards the general number of words counted.

Sentence	**Parameters**	**Sentiment score**
As shown in Fig. 5, Soluble FGL2 significantly reduced the expression of major histocompatibility complex-II, CD40, CD80, CD86, and CD83.	a=10, b=1	1.0
CD40 ligation induces or increases expression of accessory molecules such as CD80.	a=10, b=2	2.0
IFN-Î³-induced CD40 expression involves the activation of STAT-1Î± as well as the NF-ÎºB activation.	a=10, b=0	0.0
... detection of 19 individual inflammatory molecules in a single 75 Âµl brain homogenate including IL-1a, IL-1ÃŸ, TNF-a , IFN-Â¿, IL-6, IL-9, IL-10, IL-12p70, IL-12p40, IL-15, IL-17, CXCL1/keratinocyte chemoattractant , CXCL2/macrophage inflammatory protein-2 (MIP-2), CXCL9/monokine induced CXCL10/IFN-Â¿-induced protein.	a=19, b=1	0.5263158
B7 co-stimulatory molecules and CD40 regulate TNF.	a=1, b=1	10.0
STAT2 with a K390E substitution restored the pattern of the interaction between STAT2 and STAT1 observed with wild-type STAT2.	a=1, b=0	0.0

Once this proposed sentiment score was established, validating it via sentiment score distributions were necessary. Figure 12 below is a distribution of the average sentiment scores with respect to differing numbers of citations. As the number of citations increased, the quicker the average sentiment score approached 1. This is expected and poses no significant correlation as averaged will taper off as a sample size increases.

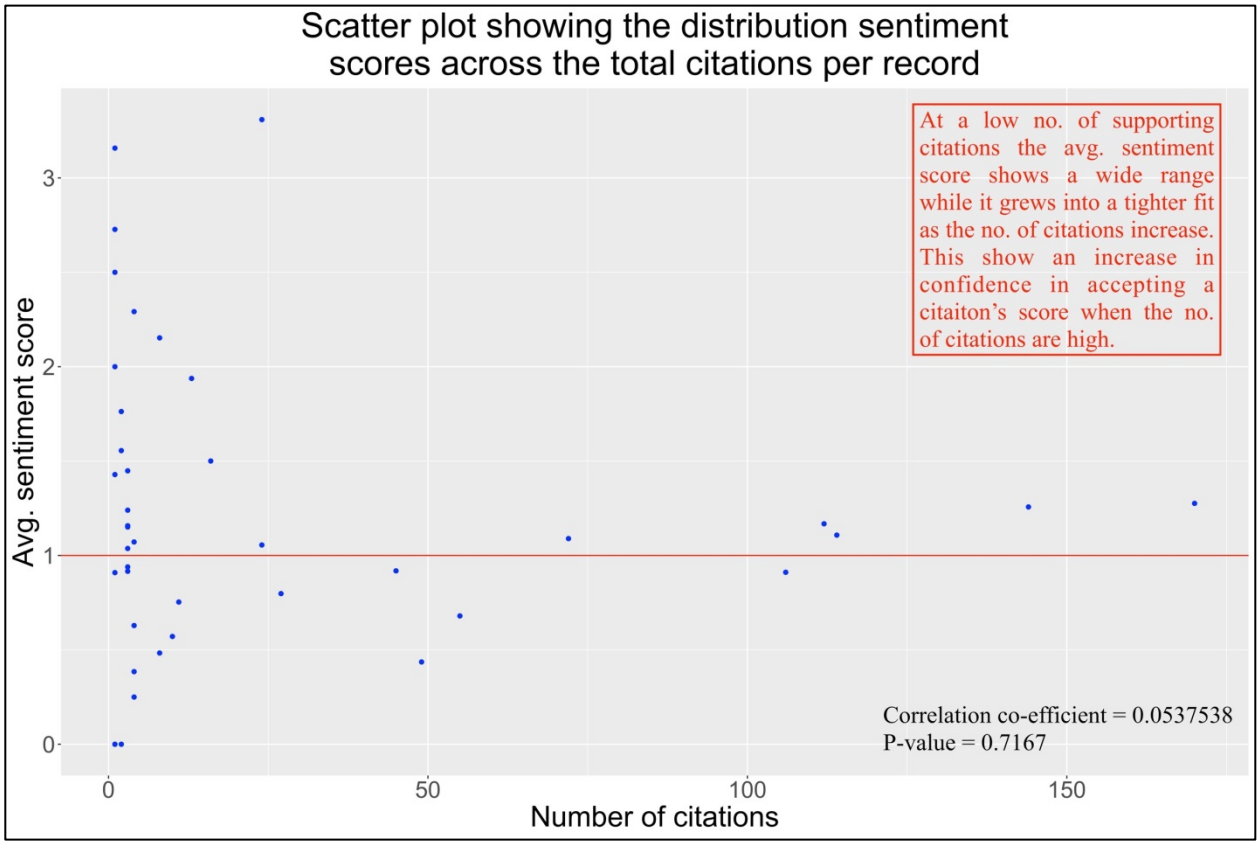

Figure 12. The distribution of sentiment scores across the total number of citations per unique source-target pairing showing that as the number of citation increase, the more likely that outliers are balances out within their respective sub-populations. This indicates that the higher the number of supporting citations is, the higher the confidence one may put into considering a readability score towards the specific search.

Figure 13 provides a different perspective by showing the sum of readability scores for different numbers of citations against a regression that represent 1 citation equal to a readability score of 1. As the number of citations increases, the sum of readability scores remains along this line with

a few outliers. There appears to be an inverse relationship compared to Figure 12 where those relationships with smaller the numbers of citations are closer remain to the ideal readability score. The purple lines represent the standard deviation derived from the data points compared to the one-to-one line, representing the range of the predicted values according to this model. Notice that there are 3 points marked 1, 2, and 3 which fall outside of this range.

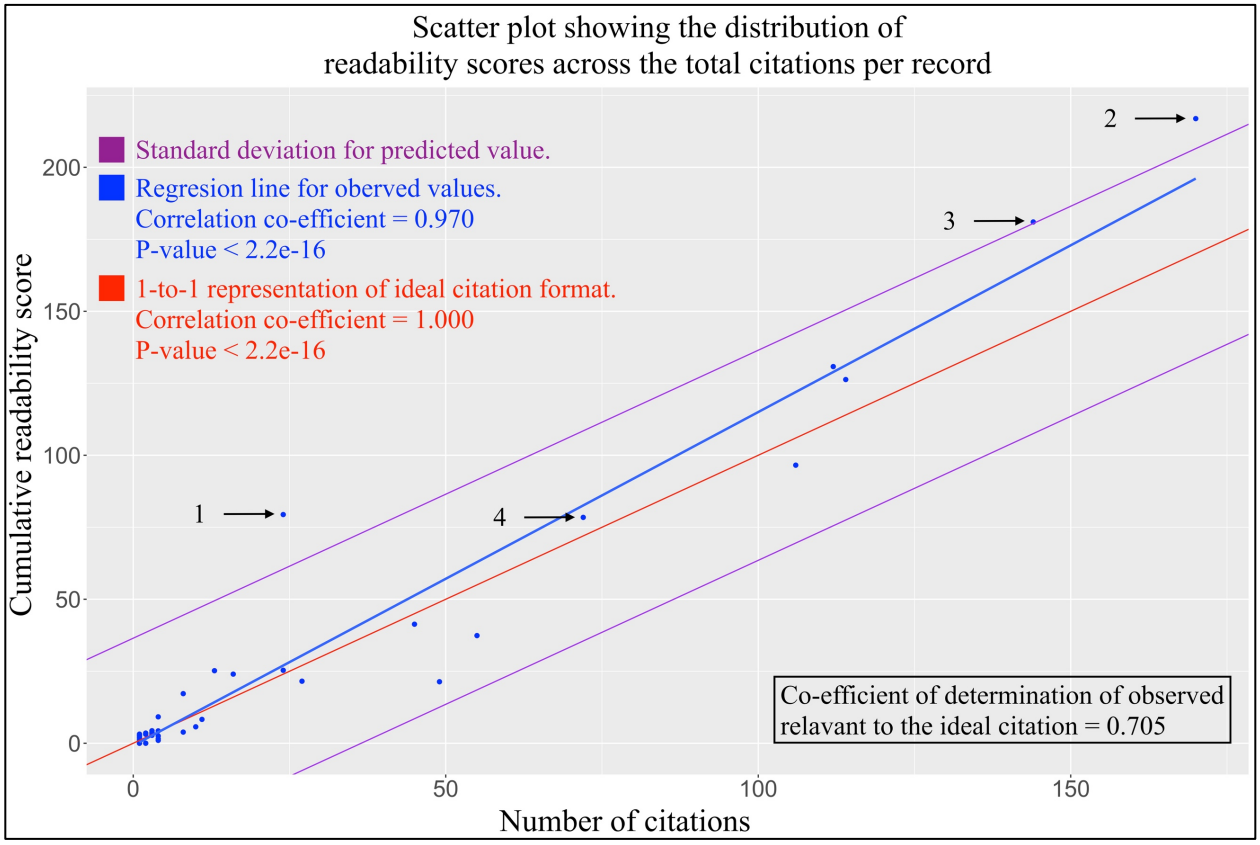

Figure 13. Scatter plot showing the distribution of sentiment scores according to the formula in figure 11, across the total number of citations per unique source-target pairing. The red diagonal line is represented by x=y where 1 citation equals a sentiment score of 1. The correlation of determination found for the observed values with respect to x=y was found to be 0.705.

Notice that Figure 13 has numbers 1 through 4 labeled at certain points in the graph. These numbers represent example source and target pairings. Number 1, 2, and 3 were chosen because

they lie outside of the predicted readability range and number 4 is a pairing that falls close to the 1-to-1 line representing the 'sweet spot' on which a citation of readability score 1.0 would fall. For each of the pairings in Table 3, there was a trend observed, the higher the readability score sum, the higher number of citations supporting them, as shown in Figure 12. The average sentiment score for this unique pairing number 1 is 79.413 with 24 supporting citations, Number 2 is 216.900 with 170 citations, number 3 has a score of 181.014 with 144 citations, and number 4 has a score of 78.433 with 72 citations supporting it.

Table 3. Table showing the unique pair and its avg. no. of words and verbs, number of supporting citations, empirical p-value, and their average sentiment score according to the labeled points found in Figure 13. The empirical p-value represents of the relationship of the observed readability score compared to the null distribution visualized in Figure 14, calculated from log transformed parameters.

No.	Source	Target	Mean. no. of words	Mean no. of verbs	No. of citations	Cumulative readability score	Empirical P-value (from log transformed params.)
1	TNF	FGL2	7.833	1.000	24	79.413	0.333
2	STAT2	STAT1	9.859	0.994	170	216.900	0.333
3	CD40	TNF	9.347	0.958	144	181.014	0.333
4	TNF	CXCL10	9.278	0.806	72	78.434	0.667

Table 4. Extension of table 3 showing the empirical p-value for every relationship formed within the network of source and targets used for this project as a product of the log transformed parameters.

Source-Target	Mean no. words (log)	Mean no. verbs (log)	No. of citations	Cumulative readability score	Empirical p-value
STAT2-STAT1	2.28836684	-0.0058997	170	216.899588	0.33333333
FGL2-STAT1	2.79320801	0.69314718	3	3.71929825	0.33333333
IFNG-STAT1	2.2512918	-0.6931472	4	1.53846154	0.66666667
CCL5-STAT1	1.60943791	0	1	0	1
FGL2-CD80	2.26868354	0	3	3.11111111	0.33333333
TNF-CD80	2.43361336	-0.4054651	45	41.3325431	0.66666667
CD40-CD80	2.30258509	0.09368548	112	130.796823	0.33333333
FGL2-CD86	2.36712361	0	3	2.81818182	0.66666667
TNF-CD86	2.49245386	-0.4519851	55	37.3810066	0.66666667
CD40-CD86	2.40643503	0.06394872	106	96.559354	0.66666667
STAT2-CCL5	2.39789527	1.09861229	1	2.72727273	0
TNF-CCL5	2.10711626	-0.8472979	49	21.3595915	0.66666667
IFNG-CCL5	2.77258872	1.38629436	1	2.5	0
CD40-CCL5	2.2512918	-0.2876821	8	17.2222222	0.66666667
CXCL2-CCL5	2.51230562	0	3	2.74853801	0.66666667
CXCL10-CCL5	2.7080502	0.11778304	8	3.8660621	0.66666667
STAT1-CCL5	2.12823171	-0.6931472	10	5.70970696	0.66666667
TNF-CSF3	1.60943791	-1.3862944	4	1	1
CCL5-CSF3	1.94591015	0	1	0	1
CXCL10-IFNG	1.09861229	0	1	0	0.33333333
CD40-TNF	2.23507921	-0.0425596	144	181.013531	0.33333333
CD200R1-TNF	2.93119375	0.40546511	4	2.51515152	0.66666667
CTSB-TNF	2.74084002	0.22314355	4	4.28571429	0.66666667
CXCL2-TNF	2.94443898	1.79175947	1	3.15789474	0
FGL2-CD40	2.15948425	0	3	3.47222222	0.33333333
TNF-CD40	2.10113437	-0.4735186	114	126.293474	0.66666667
CD200R1-CD40	2.39789527	0	1	0.90909091	0.66666667
STAT1-CD40	1.69644929	-0.7884574	11	8.28571429	0.66666667
TNF-FGL2	2.05838813	0	24	79.4134199	0.33333333
CD40-FGL2	1.38629436	0	2	0	1
TNF-CXCL2	2.47248413	0.2595112	27	21.5425982	0.33333333

CCL5-CXCL2	2.9087209	0	3	3.45238095	0.66666667
CXCL10-CXCL2	1.60943791	0	2	0	1
STAT2-CXCL10	1.94591015	0	2	3.11111111	0.33333333
TNF-CXCL10	2.22762205	-0.2162231	72	78.4338548	0.66666667
IFNG-CXCL10	1.79175947	0	1	0	1
CD80-CXCL10	2.07944154	0	1	0	1
CD40-CXCL10	1.178655	-0.2876821	4	9.16666667	0
CXCL2-CXCL10	2.23359222	0.28768207	3	4.3452381	0.33333333
STAT1-CXCL10	1.7122953	-1.0986123	24	25.3333333	0.66666667
CD86-CXCL10	1.94591015	0	1	0	1
TNF-CTSB	2.94443898	1.38629436	2	3.52490421	0
CXCL10-CTSB	1.09861229	0	1	0	1
TNF-NFKB1	2.36320971	0.40546511	16	24.005558	0.33333333
TNF-STAT2	1.60943791	0	1	2	0.33333333
IFNG-STAT2	1.94591015	0	1	1.42857143	0.33333333
CXCL10-STAT2	2.19722458	0	1	0	1

The trend observed here is that the higher the number of citations supporting any given relationship, the higher the calculated readability score as shown in Figure 11. However, while looking at sums instead of averages, we see that these points (except the point designated number 4) are very verb-rich. This is because the score does not follow a linear trend, the higher the number of citations is, the higher the probability of including verb-rich statements, therefore boosting the sum of scores higher than the rate of lesser supported relationships.

To conclude this project's analyses the sentiment scoring was performed on categories B, C, and D to find what the average sentiment score was in each. Since the remaining categories do not have a 100% chance of having a verbatim source and target, instead data set was altered to take into consideration all possible pairings within every citation. This causes the number of citations to expand within each group but not necessarily using all for the analysis. This can be seen in Table 5 below where each category is shown with their sentiment score statistics

compared to one another. It was found that for categories A, B, C, and D, the readability score remained around 1.000 as 1.058, 1.012, 1.069, 1.194 respectively, even after expanding the sample size within.

Table 5. Table showing the average sentiment scores for each category of this project's dataset. The average sentiment score represents the average sentiment score found for every used citation in each category using the proposed formula in Figure 11.

Category	Avg. sentiment score	Original no. of citations	Expanded no. of citations	Portions of the expanded used
A	1.058	1123	3582	3582/3582 (100.00%)
B	1.012	1274	1579	561/1579 (42.50%)
C	1.069	1560	2470	1717/2470 (69.51%)
D	1.194	1013	1078	143/1078 (13.27%)

After expanding the datasets, Category A was once again used to visualize the distribution of readability scores of the total 3582 citations. In Figure 14, the data was observed to be skewed left which indicated the mean was lower than the median of this category with a larger standard deviation, and as a log transformation in figure 15. The 95% confidence interval was displayed and observed to fall close the mean readability score of 1.058. This is to say that 95% of the readability scores within this category remains close the ideal citation format, leading more confidence it the application of the readability scoring formula to larger datasets.

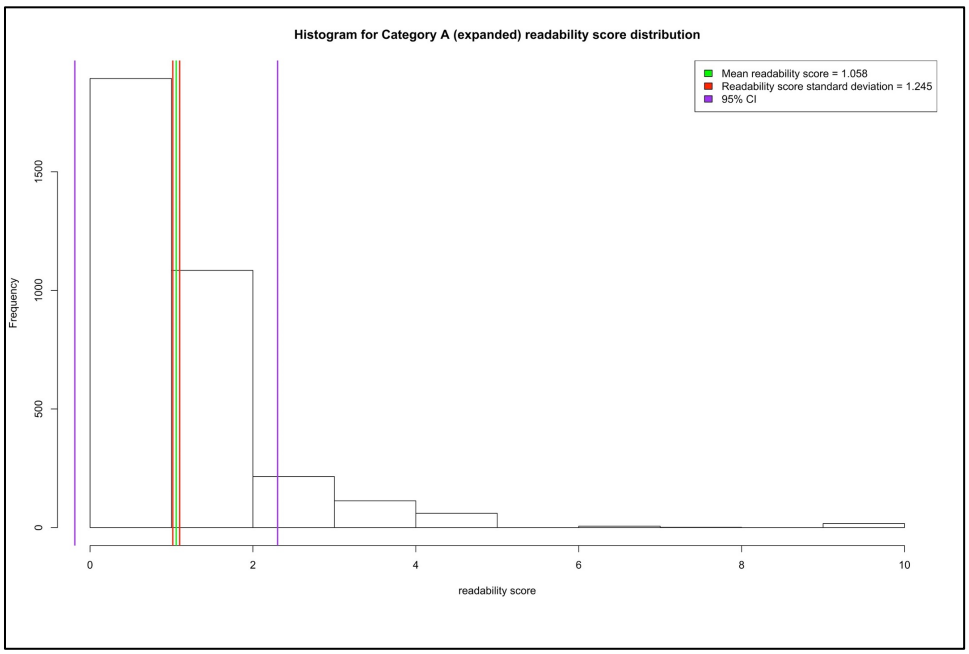

Figure 14. Distribution of readability score within the expanded version of category A which includes 3582 total citations. The mean readability score was calculated to be 1.058 while the standard deviation is 1.245.

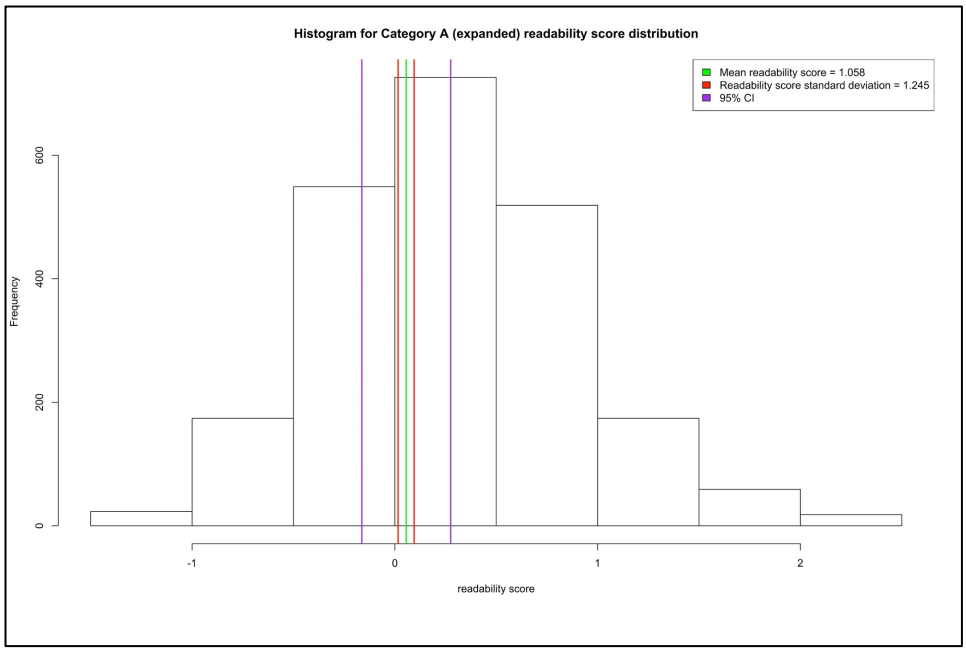

Figure 15. Log transformation of the observed readability score within Category A.

Discussion

The current project began by establishing an average citation structure that consists of 10 words in between a given source-target pair with 1 of those words being a verb. This information is valuable as it defines a standard way that the scientific community has found to be the most effective. This, however, was observed (Figure 8) as having one exception where there is one word separating the pair and that word is a verb. In many cases this can be seen as a good example of a direct relationship. Beyond this exception, as the number of words and verbs vary the readability scores begins to change, either above or below 1. This indicates what the nature of this variation from 1 may be. As the number of words increases while keeping the verb count steady, the score rapidly increases as seen in Table 2, this was to say that with higher number of words but proportionally lower number of verbs, the citation was most often not an understandable citation. Once the verb count began to increase the score did as well towards 10 as the number of verbs became proportional to the number of words. These would be interpreted as verb-rich statements which do not necessarily form a readable citation. It is important to note that this formula is restricted to being used in the case of 1 or more words. Zero represents the inability to discern whether a relationship is formed when two words are directly beside each other. This is a statistic that can be accompanied by what is observed in Figure 12, however, so that one may see that a pairing must be supported by multiple citations to be confidently considered. Figure 12 demonstrates that the higher the number of citations that are attributed to a pairing, the closer its average sentiment score comes to 1. This simply says that a higher number of citations gives rise to higher confidence that the source-target pairing in fact indicates that the source and target have an actual relationship to one another.

Now, with respect to the readability scoring formula proposed in Figure 11, consider Elsevier's PS software as a control. It was found that 85.7% of those citations that were included in category A fell within the threshold indicated by the standard deviation ellipse (Friendly et al., 2013) as shown in Figure 10. What this says is that 14.3% of the citations seemed to much longer citations than the observed structure as any point outside of the data ellipse represented pairing that had much higher number of words and verbs having occurred within the top right quadrant in Figure 10.

Within Table 3, there are 4 observed unique pairings that have been chosen to explore because of their abnormally high cumulative readability score. These are indicated in Figure 13 as numbers 1 through 4, and their statistical attributes displayed in Table 3. For each of these relationships, the p-value was calculated and found to be 0.456, 0.595, 0.509, and 0.444 respectively. This was followed by converting the calculation parameter to log transformed values, resulting in gin 0.333, 0.333, 0.333, and 0.667 respectively. What this tells us about the distribution of readability scores is that for each of these relationships, we fail to reject the null distribution, meaning that there is no statistical significance to the relationships between verb and word count for these found in table 3. The project moves to expand this table log transforming the parameters and performing an empirical p-value calculation on each of the relationships in the network in order to find true sample distribution. It was found that 20.84% of the relationship having an empirical p-value from the log transformed parameters as 1.000 while the rest were within range of 0.000-0.700. Number 1 appears very high above the 1-to-1 line that represents the average instance. This is because the sum of that pairing's sentiment score is a lot higher for the number of citations which supports it, 24. This indicates that even though there are a low number of citations supporting it, there are a lot of verb-rich statements included which raises the

cumulative readability score higher than expected. The expected meaning that for each citation the average would be a score of 1. Having a such a high score also indicates a citation that is verb-rich, which is more indicative of being readable and forming a relationship between the source and target. Number 2 shows this as well with its number of supporting citations being 170, in a perfect situation along the lines of the established average citation, all scores are 1.0. This would mean a cumulative score of 170. This does not mean all its citations are very rich but rather a lot of them are, not just every citation having a score of approximately 2. Number 3 shows the same, but it falls right above the threshold of predicted values as seen in Figure 12. Number 4 on the other hand is an interesting observation compared to numbers 1, 2, and 3. Even through the number of supporting citations are 72, its cumulative sentiment score hovers around the 1-to-1 diagonal. This indicated that on average this unique pairing must have has a sentiment score hovering around 1.0 for each citation. This is observed as a normal performer according to the observed average citation format of having 1 verb and 10 words. Additional support to this is the correlation co-efficient and its p-value for Figures 12 and 13. To have trust that the red lines that indicate the average citation format these statistical attributes were calculated. In Figure 12 the red line indicated the position of having a sentiment score of 1.0. The correlation coefficient of this graph was found to be 0.0538 with a p-value of 0.7. What this indicates is that the average sentiment score does not grow with the number of citations in a way to be established as having a positive correlation but the p-value being higher than 0.05 indicates that this is not significant. What this graph does in fact show however is that no matter how higher the average sentiment score may be, the more supporting citations occurs then the more outliers are balanced out to the average of 1.0. The opposite is observed in Figure 13 where the correlation co-efficient is 0.970 which is indicative of a positively correlating relationship, and a p-value of less than 2.2e-16.

This extremely low p-value indicates that the observed correlation co-efficient is significant. In addition to this, the co-efficient of determination of the observed data with respect to an ideal citation represented by the 1-to-1 diagonal line in Figure 13 was found to be 0.705. This represents that there is a 70.50% variation in the observed values with respect to the average citation format found earlier.

To conduct this correlation testing, the Pearson test was used over either Spearman or Kendall correlation tests. The Pearson test was used because the 4 assumptions (Obilor & Amadi, 2018) that it follows fit the data set better than those of Spearman and Kendall tests. Assumption 1 is that the data used is a ratio, this can be said for the word-to-verb ratio used to calculate the readability scoring. Assumption 2 is that it works with linear data, and this is supported the fact that with an increasing number of citations there is an increasing number of averages and cumulative readability scores. Assumption 3 states that the outliers found within the data can perturb the fit of the data point. Lastly, assumption 4 is that the data is observed to be normally distributed. The Spearman test is slightly different in that it relies on the same assumptions as the Pearson test with the exception that the data need not be normally distributed (Hauke & Kossowski, 2011). The Kendal test was also not used because that test works by looking for dependencies rather than correlation. Dependency in statistics means to have the value of one variable assigned based on the value of another (Ye et al., 2015).

The last step of this project applied the proposed formula to categories B, C and D, the results of which can be seen in Table 5. This table includes an average readability score for each category, and this was achieved by creating a library of all possible source and target keys and using it to form every unique pairing possible between any 2 of the possible relationships. If a citation had only one verbatim appearance from the library, then it was not counted towards the

number of supporting citations and if one citation had multiple pairings then it was counted according to the number of pairings. As expected, category A utilized 100% of its citations to produce an average sentiment score of 1.06. Category D utilized the least number of its total citations and had the higher readability scoring. However, it was still with an acceptable margin to be counted as a standard citation format. Categories B and C utilized 42.50% and 69.51% respectively. It is not fair to say it makes sense as to which category having the higher percentage could have been different considering there are working within the same terms, only one verbatim appearance of at least either the source or the target. Their readability scores do remain around 1.0 however, which supports the notion that an average of 10 words and 1 verb is the average composition of a citation between a source and a target is possible. This is especially true since the unique pairings were not used but instead any possible combinations seen within citations were used, thereby increasing the sample size within each category.

Upon expanding the dataset of each category, visualization of the distribution for category A's expanded form was done, as observed in Figure 14. This distribution displays a left skew distribution which indicates that the mean is smaller than the median for this sample, and 95% confidence interval shows 95% of the data aggregates close to the mean of 1.058. Category A alone is enough to indicate that this formula may be applicable to larger datasets because of the increased fold number of supporting citations. The initial number of pairings for this was 48, being supported by 1123 citations. After expanding this grew to 86 pairings being supported by 3582 citations. The project can conclude that the proposed formula has wider application based on the readability score observed.

Conclusion

Although the project scope was very limited considering the dataset used was approximately a quarter of the total 4970 citations produced by PS within the establish relational network described in Figure 1. It was important to establish this environment to give the results some a certain level of confidence by knowing the exact condition of this analysis. The project was successful in gaining more insight into trend of citation structures found within life science literature that is used for different text mining methods. The average statement found within the dataset was found to consisted of 1 verb and 10 words between a source and target, which was later justified by the observed distribution of the readability score and 95% CI in Figure 14.

The statistical information derived poses useful for future work on expanding the scope of the project to incorporating synonyms when text mining as it sets the foundation for moving onto including alternative forms instead of just considering verbatim occurrences. This project opens future work to allow the expansion to include adjectives in the list of verb types as an important improvement as it was observed during this project that there are more parts of speech that may be indicative of a relationship formed between a source and target, as well as increasing readability by having more types of linkages between the pair. This is because adjectives were noticed to sometimes show a relationship between source and target.

Proposed next steps for this project would be to either increase the sample sizing of each category to be the same to create a leveled playing field or normalize the sentiment scoring according to their percentage of citations that make up categories B, C and D in order to have a more consistent sample size for each category for comparison of performance with the readability score.

References

Adams, E. J., Nguyen, A. T., & Cowan, N. (2018). Theories of Working Memory: Differences in Definition, Degree of Modularity, Role of Attention, and Purpose. *Language, Speech, and Hearing Services in Schools*, *49*(3), 340–355. https://doi.org/10.1044/2018_LSHSS-17-0114

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In M. W. Berry, A. Mohamed, & B. W. Yap (Eds.), *Supervised and Unsupervised Learning for Data Science* (pp. 3–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-22475-2_1

Anjali, M K, and Babu P Anto. "Ambiguities in Natural Language Processing." *International Journal of Innovation REsearch in Computer and Communication Engineering*, vol. 2, no. 5, Oct. 2014, pp. 392–294., doi:http://rroij.com/open-access/ambiguities-in-natural-language-processing.pdf.

Apache OpenNLP Developer Documentation. (n.d.). Retrieved June 23, 2021, from https://opennlp.apache.org/docs/1.9.2/manual/opennlp.html

Artetxe, M., & Schwenk, H. (2019). Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, *7*, 597–610. https://doi.org/10.1162/tacl_a_00288

Datacube Resource Homepage. (n.d.). Retrieved June 23, 2021, from http://datacube.wu.ac.at/

Deveci, T. (2019). *Sentence Length in Education Research Articles: A Comparison between Anglophone and Turkish Authors*.

Do values in a numeric vector fall in specified range? — Between. (n.d.). Retrieved June 24, 2021, from https://dplyr.tidyverse.org/reference/between.html

Friendly, M., Monette, G., & Fox, J. (2013). Elliptical Insights: Understanding Statistical Methods through Elliptical Geometry. *Statistical Science*, *28*(1). https://doi.org/10.1214/12-STS402

Hauke, J., & Kossowski, T. (2011). Comparison of Values of Pearson's and Spearman's Correlation Coefficients on the Same Sets of Data. *QUAGEO*, *30*(2), 87–93. https://doi.org/10.2478/v10117-011-0021-1

Kher, A., Johnson, S., & Griffith, R. (2017). Readability Assessment of Online Patient Education Material on Congestive Heart Failure. *Advances in Preventive Medicine*, *2017*, 9780317. https://doi.org/10.1155/2017/9780317

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2017). *Natural Language Processing: State of The Art, Current Trends and Challenges*.

Maxent_POS_Tag_Annotator function—RDocumentation. (n.d.). Retrieved June 24, 2021, from https://www.rdocumentation.org/packages/openNLP/versions/0.2-7/topics/Maxent_POS_Tag_Annotator

Morris, M. C., Lyman, C. A., Richman, S., Cao, H. B., Cheadle, C., & Broderick, G. (2020). Predicting the Immune Response to Repurposed Drugs in Coronavirus-induced Cytokine Storm. *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, 458–465. https://doi.org/10.1109/BIBE50027.2020.00080

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. *Journal of the American Medical Informatics Association : JAMIA*, *18*(5), 544–551. https://doi.org/10.1136/amiajnl-2011-000464

Nikitin, A., Egorov, S., Daraselia, N., & Mazo, I. (2003). Pathway studio—The analysis and navigation of molecular networks. *Bioinformatics*, *19*(16), 2155–2157. https://doi.org/10.1093/bioinformatics/btg290

Obilor, E. I., & Amadi, E. (2018). *Test for Significance of Pearson's Correlation Coefficient ( )*.

Perez, O. D., Swindell, H. W., Herndon, C. L., Noback, P. C., Trofa, D. P., & Vosseller, J. T. (2020). Assessing the Readability of Online Information About Achilles Tendon Ruptures. *Foot & Ankle Specialist*, *13*(6), 470–477. https://doi.org/10.1177/1938640019888058

Plavén-Sigray, P., Matheson, G. J., Schiffler, B. C., & Thompson, W. H. (n.d.). The readability of scientific texts is decreasing over time. *ELife*, *6*, e27725. https://doi.org/10.7554/eLife.27725

Roy, B., & Purkayastha, B. S. (n.d.). *A Study on Different Part of Speech (POS) Tagging Approaches in Assamese Language*. *5*(3), 5.

Safeer, R. S., & Keenan, J. (2005). Health Literacy: The Gap Between Physicians and Patients. *American Family Physician*, *72*(3), 463–468.

Santorini, B. (n.d.). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. 37.

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Saussure, F., & De, 1857-1913. (n.d.). *Course in General Linguistics*. 257.

Sharp, S. (n.d.). *Pathway Studio- Assisted Biological Research*. 43.

Stoltz, D. S. (n.d.). *Becoming A Dominant Misinterpreted Source: The Case of Ferdinand De Saussure in Cultural Sociology*. 34.

Tokenization as the initial phase in NLP. (n.d.). https://doi.org/10.3115/992424.992434

Ye, L., Zhou, J., Zeng, X., & Tayyab, M. (2015). Hydrological Mann-Kendal Multivariate Trends Analysis in the Upper Yangtze River Basin. *Journal of Geoscience and Environment Protection*, *03*(10), 34–39. https://doi.org/10.4236/gep.2015.310006

Rochester Institute of Technology
Thomas H. Gosnell School of Life Sciences
Bioinformatics Program

To: Head, Thomas H. Gosnell School of Life Sciences

The undersigned state that Jeremy Jackson, a candidate for the Master of Science degree in Bioinformatics, has submitted his thesis and has satisfactorily defended it.

This completes the requirements for the Master of Science degree in Bioinformatics at Rochester Institute of Technology.

Thesis committee members:

Name	**Date**
Gary R. Skuse, Ph.D. Thesis Advisor	
Gordon Broderick, Ph.D.	
Matthew Morris, Ph.D.	

Feng Cui, Ph.D.
Director of Bioinformatics MS Program

475-4115 (voice)
fxcsbi@rit.edu